Master Thesis

Unblinded sample size re-estimation for diagnostic accuracy studies in an unpaired comparative design

Denise Köster, Antonia Zapf, Annika Hoyer

GMDS-TMF conference, September 27, 2021

Universitätsklinikum Hamburg-Eppendorf

Overview

1. Background
2. Methods for the comparison of diagnostic tests
   i. Statistical aspects
   ii. Simulation Study
   iii. Example
3. Summary

Confirmatory Diagnostic Accuracy Studies

- Diagnostic accuracy
- Co-primary endpoints

index test	gold/reference standard	
	diseased	non-diseased
positive		
negative		

sensitivity

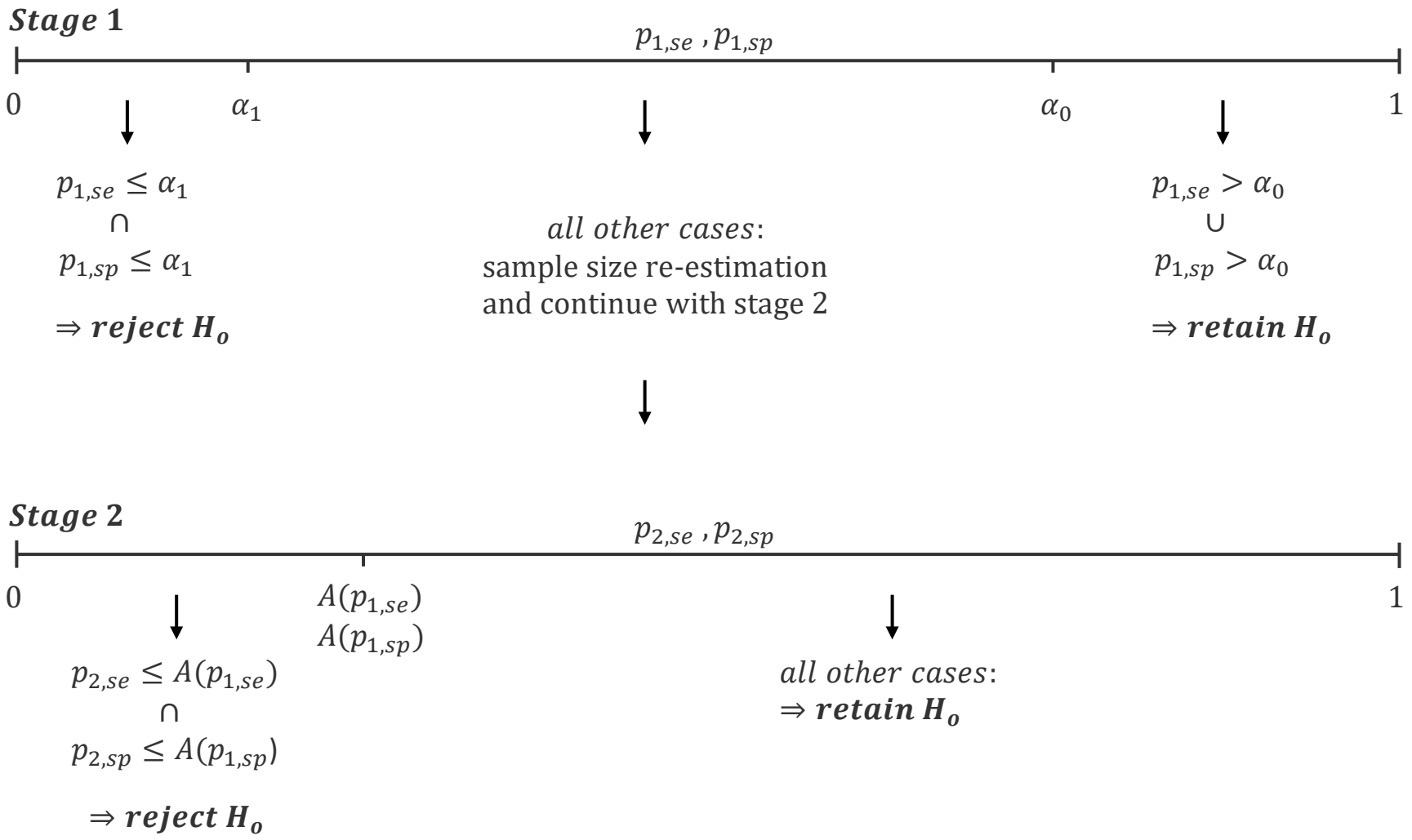
$$se = \frac{TP}{TP+FN}$$

specificity

$$sp = \frac{TN}{TN+FP}$$

Adaptive Designs in Diagnostic Accuracy Studies

- **Change design aspects during an ongoing study**
  - Sample size re-calculation and patient recruitment
- **Blinded vs. unblinded interim analysis**
  - Estimation of, e.g. prevalence vs. sensitivity and specificity
  - Type I error is not affected vs. needs to be adjusted
  - Sample size re-calculation on basis of the interim data
- **Ethical, moral, time and financial reasons**


Stage 1
$p_{1,se}$, $p_{1,sp}$
0
α_1
α_0
1
$p_{1,se} \leq \alpha_1$
∩
$p_{1,sp} \leq \alpha_1$
⇒ reject H_o
all other cases:
sample size re-estimation
and continue with stage 2
$p_{1,se} > \alpha_0$
∪
$p_{1,sp} > \alpha_0$
⇒ retain H_o
Stage 2
$p_{2,se}$, $p_{2,sp}$
0
$A(p_{1,se})$
$A(p_{1,sp})$
1
$p_{2,se} \leq A(p_{1,se})$
∩
$p_{2,sp} \leq A(p_{1,sp})$
⇒ reject H_o
all other cases:
⇒ retain H_o

Hypotheses

Aim: Comparison of an experimental test with a comparator test in two possible hypotheses settings

Setting 1: Prove that the sensitivity (se_E) and specificity (sp_E) of the experimental test are different from the sensitivity (se_C) and specificity (sp_C) of the comparator test

$$H_{0,se}: se_E \leq se_C \quad \cup \quad H_{0,sp}: sp_E \leq sp_C$$
$$H_{1,se}: se_E > se_C \quad \cap \quad H_{1,sp}: sp_E > sp_C$$

Hypotheses

Setting 2: Prove that the sensitivity (se_E) of the experimental is different from the sensitivity (se_C) of the comparator test, and the specificity (sp_E) of the experimental test is at least as good as the specificity (sp_C) of the comparator test within a non-inferiority margin (δ_{sp})

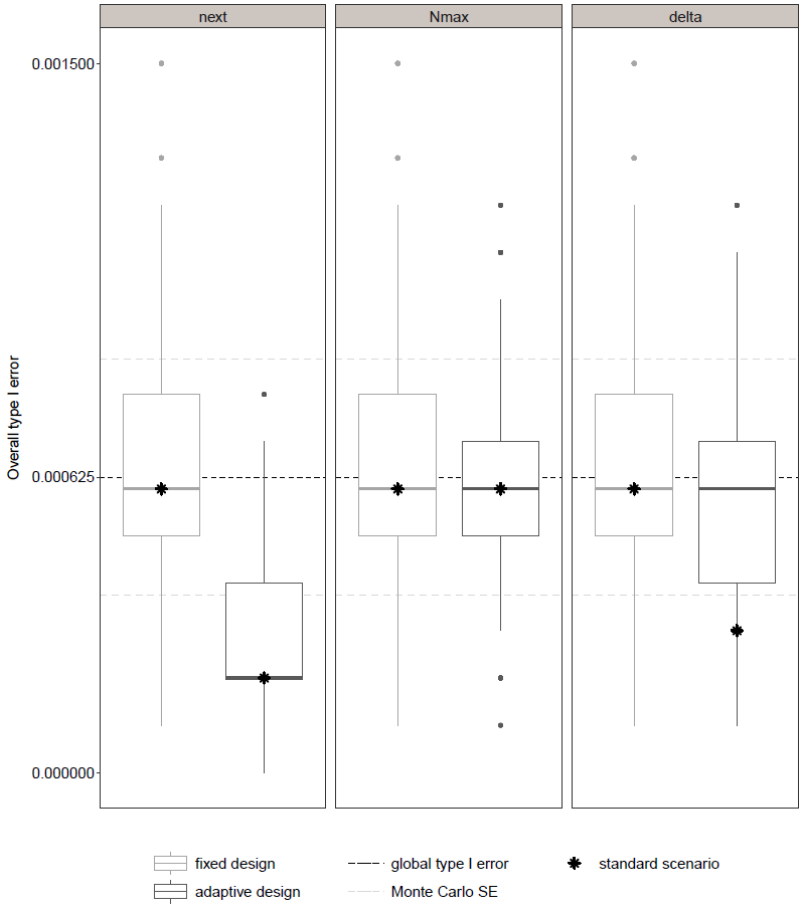
$$H_{0,se}: se_c = se_e \cup H_{0,sp}: sp_c - sp_e \geq \delta_{sp}$$
$$H_{1,se}: se_c \neq se_e \cap H_{1,sp}: sp_c - sp_e < \delta_{sp}$$

Fundamentals

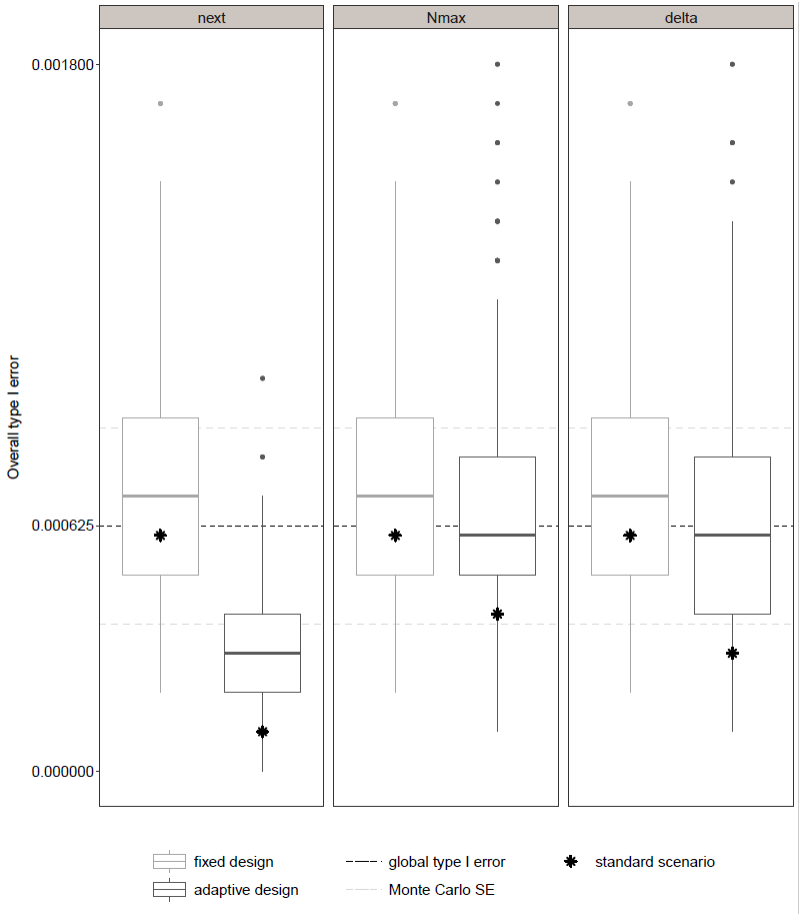
Aim: Comparison of the adaptive and the fixed design regarding the type-one error α and the statistical power 1 – β for both hypotheses settings

- Implementation in R
- Variety of scenarios representing realistic constellations
  - $\alpha = 2.5\%$ (one-sided), $1 - \beta = 80\%$
- Arbitrarily chosen standard scenarios as a reference setting
- Pre-specified maximum sample size for each scenario
- Consideration of three options at the interim analysis (next, Nmax, delta)

Results – Type I error rate

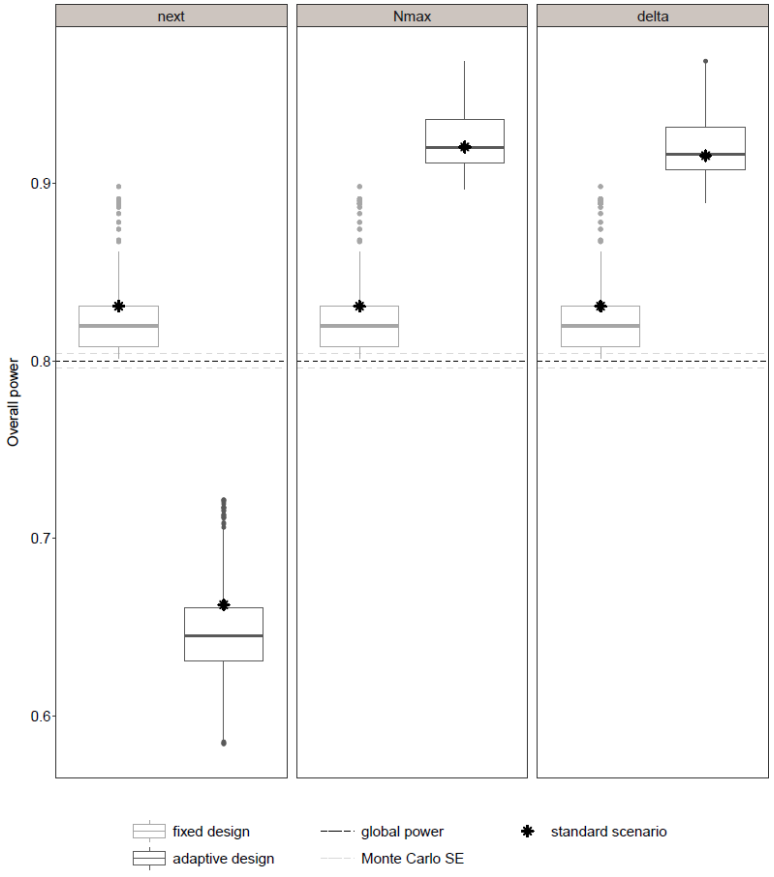


Hypotheses setting 1

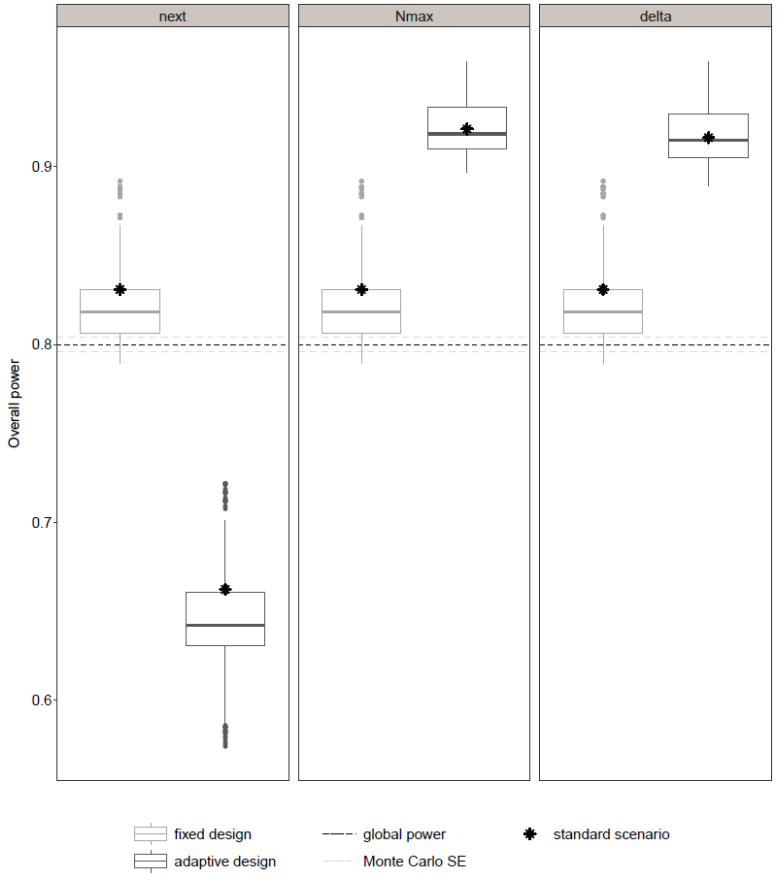


Hypotheses setting 2

Results – Power



Hypotheses setting 1



Hypotheses setting 2

European Journal of Cancer Prevention 2012, 21:460–466

Clinical evaluation of an autofluorescence diagnostic device for oral cancer detection: a prospective randomized diagnostic study

Majeed Rana[a], Antonia Zapf[b], Marco Kuehle[a], Nils-Claudius Gellrich[a] and André M. Eckardt[a]

- **Gold standard:**
  biopsy
- **Comparator:**
  Conventional oral examination (COE) using whitelight
- **Experimental:**
  COE with an autofluorescence visualization device (VELscope) and whitelight

- **Previous knowledge:**

Statistical analysis

The sample size for the study was planned using the data of a pilot study ($n = 30$). In this pilot study, the white light examination showed a sensitivity of 50% and a specificity of 100% and for white light plus VELscope the result showed a sensitivity of 100% and a specificity of 96%. The aim of

Results of the Example

Aim: demonstrate that the sensitivity is higher and the specificity is not relevant lower

Assumptions: $se_C = 50\%$, $sp_C = 100\%$, $se_E = 100\%$, $sp_E = 96\%$, $\delta_{sp} = 20\%$,
→ Needed sample size per diagnostic test N = 150 ($\alpha = 5\%$, $1 - \beta = 90\%$, $\pi = 10\%$)

Results:

The results of the evaluation of the diagnostic accuracy are shown in Table 4. As expected, the additional use of the VELscope led to a higher sensitivity (100% instead of 17%), but to lower specificity (74% instead of 97%) (Figs 1–3).

Results of the Example for the Adaptive design with option delta

1. Initial sample size per test: 170
   → Use optimal sample size calculation with the prevalence to reach the desired power of 90%
2. Calculate the maximum sample size: 256
3. Number of simulation runs: 10,000
4. Recruitment of half of the initial sample size per test: 85

Results of the Example for the Adaptive design with option delta

6. Interim analysis

Early stop for

- Efficacy: 0.13%
- Futility: 46.42%
- sample size re-calculation: 32.44%
- Transition stop: 50.41%
- Maximum sample size used: 1.67%

96,83%

7. Final analysis: 3.04%

- Efficacy: 0.18%
- Futility: 2.86%

→Overall power:

- Fixed design: 0.48%
- Adaptive design: 0.31%

$Prevalence$ 7.1% (5%)	**COE + whitelight**	**COE + VELscope + whitelight**
$mean\ \widehat{sensitivity}$	28.32% (17%)	82.75% (100%)
$mean\ \widehat{specificity}$	95.78% (97%)	73.36% (74%)

Estimates of the example study in brackets

Keypoints

- Increased complexity of diagnostic studies due to two co-primary endpoints
- Adaptive design: Prove that, e.g. sensitivity and specificity of the experimental test are different from the sensitivity and specificity of the comparator test
- Allowing for early stopping for efficacy or futility or sample size re-estimation while accounting for type-one error
- Adaptive designs are feasible and helpful in confirmatory diagnostic accuracy studies in an unpaired comparative design.

References

- Pepe, M. S. (2003). The statistical evaluation of medical tests for classification and prediction (Vol. 28). Oxford Univ. Press.
- Wassmer G, Brannath W (2016). Group sequential and confirmatory adaptive designs in clinical trials. Heidelberg, Springer.
- Rana, M., Zapf, A., Kuehle, M., Gellrich, N.-C., & Eckardt, A. M. (2012). Clinical evaluation of an autofluorescence diagnostic device for oral cancer detection: a prospective randomized diagnostic study. European journal of cancer prevention: the official journal of the European Cancer Prevention Organisation (ECP), 21 (5), 460–466.
- Stark M, Zapf A (2019). Sample size calculation and reestimation based on the prevalence in a single-arm confirmatory diagnostic accuracy study. Stat Methods Med Res, under revision.

UKE
HAMBURG

Do you have any questions?